STATUS OF THE MACHINE LEARNING EFFORTS AT THE INTERNATIONAL DATA CENTRE OF THE CTBTO

Ronan J. Le Bras[1], Sheila Vaidya[2], Jeff Schneider[3], Stuart Russell[4], and Nimar Arora[4]

Comprehensive Nuclear-Test-Ban Treaty Organization[1], Lawrence Livermore National Laboratories[2],
Carnegie Mellon University[3], and the University of California at Berkeley[4]

ABSTRACT

Machine learning projects were conceived in March 2009 as part of the International Scientific Studies Project initiative at the Provisional Technical Secretariat of the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) and initiated a few months later. Some of the projects are intended to aim at short to medium term operational applications. These include the identification of seismic and hydroacoustic phase names using a large number of features extracted from the waveforms, and the labeling of automatic events as real or false depending again on a large number of features from the automatic events. Concrete research results using International Data Centre (IDC) data are available for these two sets of projects. Seismic phase identification is shown to have the potential to improve its accuracy by 23 %, and the software developed for the project on false events identification has been tested at the IDC and shown to correctly label 80% of the false alarms. Some projects are aimed at the longer term. This is the case of a Bayesian approach to the automatic seismic network processing problem and a distributed database approach to the waveform cross-correlation problem. The first project is well under way and has surpassed the current operational system by 14 % in accuracy for the same false alarm rate. The second has shown the potential of distributed systems to solve efficiency issues.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **SEP 2010**	2. REPORT TYPE	3. DATES COVERED **00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE **Status of the Machine Learning Efforts at the International Data Centre of the CTBTO**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Comprehensive Nuclear-Test-Ban Treaty Organization,Vienna International Centre,P.O. Box 1200,1400 Vienna, Austria,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Published in Proceedings of the 2010 Monitoring Research Review - Ground-Based Nuclear Explosion Monitoring Technologies, 21-23 September 2010, Orlando, FL. Volume II. Sponsored by the Air Force Research Laboratory (AFRL) and the National Nuclear Security Administration (NNSA). U.S. Government or Federal Rights License

14. ABSTRACT

Machine learning projects were conceived in March 2009 as part of the International Scientific Studies Project initiative at the Provisional Technical Secretariat of the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) and initiated a few months later. Some of the projects are intended to aim at short to medium term operational applications. These include the identification of seismic and hydroacoustic phase names using a large number of features extracted from the waveforms, and the labeling of automatic events as real or false depending again on a large number of features from the automatic events. Concrete research results using International Data Centre (IDC) data are available for these two sets of projects. Seismic phase identification is shown to have the potential to improve its accuracy by 23 %, and the software developed for the project on false events identification has been tested at the IDC and shown to correctly label 80% of the false alarms. Some projects are aimed at the longer term. This is the case of a Bayesian approach to the automatic seismic network processing problem and a distributed database approach to the waveform cross-correlation problem. The first project is well under way and has surpassed the current operational system by 14 % in accuracy for the same false alarm rate. The second has shown the potential of distributed systems to solve efficiency issues.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT **Same as Report (SAR)**	18. NUMBER OF PAGES **9**	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**			

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

OBJECTIVES

The objective of this project is to investigate the application of Machine Learning techniques to the processing of waveform data at the IDC of the CTBTO. The initial motivation was that a large database of waveform and parametric data has been accumulated in the last ten years and this database is currently hardly exploited in terms of its potential to improve on existing models and processing techniques. Analyst knowledge has been captured in the archive but is not used in the current automatic system which tends to treat an incoming event as if it is seeing it for the first time without the benefit of recalling information pertaining to events in similar locations that have previously been recorded. In addition, methods that were too computationally intensive just a few years ago could now be used on a real-time basis and inserted into the IDC processing pipeline.

The ISS09 project initiated by the CTBTO in 2008 included a *Data Mining/Machine Learning* component, which was a new area of investigation for the organization. A workshop in March 2009, in preparation for the June 2009 conference, identified a number of objectives that could be fulfilled in the near term with a high probability of success and others that would take a longer effort to bear fruit in the operational system.

- The projects with operational short term goals included:
    - False Events Identification (FEI) using Support Vector Machine (SVM) methods (Mackey et al., 2009)
    - Hydroacoustic and Seismic phase identification (Tuma M. and Igel C., 2009; Schneider et al., 2010)

- The projects with operational long term goals included:
    - Vertically Integrated Seismic Analysis (VISA) detection, association, and location (Arora et al., 2009a, 2009b)
    - Distributed database approach to the waveform cross-correlation problem

Figure 1 shows the areas of impact of four of these projects within the context of the IDC waveform processing.

RESEARCH ACCOMPLISHED

The various short-term and long-term projects tackled in the Machine Learning area during the last year have led to a number of publications illustrating the benefits that can be obtained from applying concepts in that field to the problem of processing of seismic and hydroacoustic data at the IDC.

False Events Identification

The first short-term project which was tackled was the false event identification (FEI) in the SEL3 bulletin based on features of the detections and associations that comprise each hypothesized event. The problem was formulated as a classification problem of labeling SEL3 events as either true or false based on a large set of features. Several variants of the SVM methods as well as the naive Bayes methods were tested in order to identify the most promising approach.

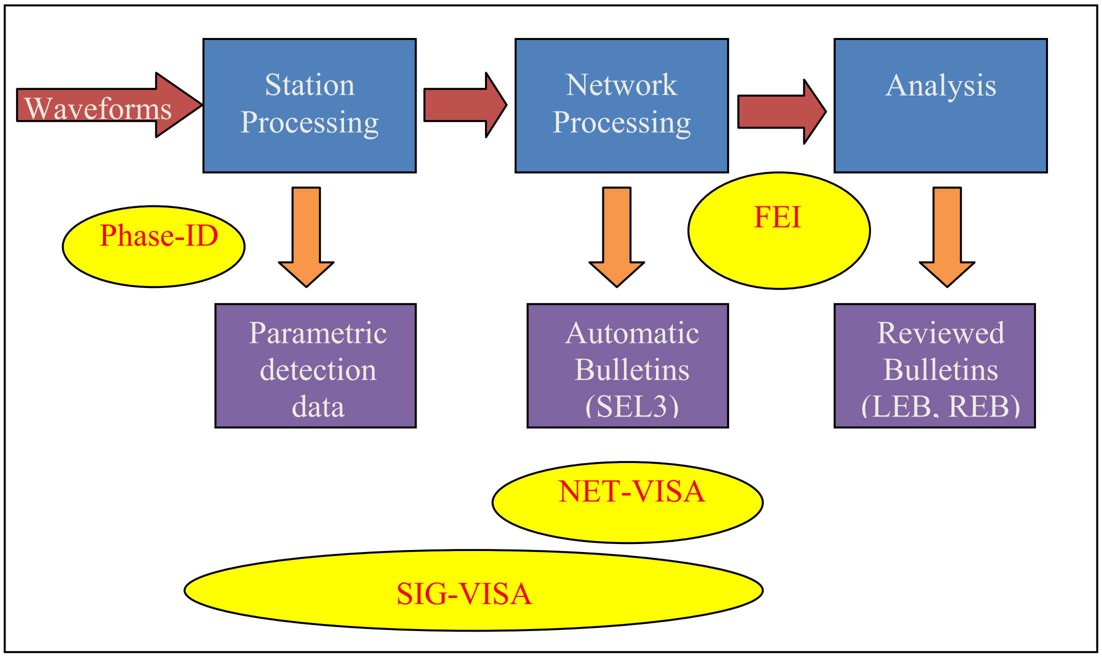
Waveforms
Station Processing
Network Processing
Analysis
Phase-ID
FEI
Parametric detection data
Automatic Bulletins (SEL3)
Reviewed Bulletins (LEB, REB)
NET-VISA
SIG-VISA

Figure 1. Simplified context diagram situating the various projects undertaken in the Machine Learning area and applied to the processing of seismic and hydroacoustic data at the IDC.

The classifiers were trained on parametric data which included both automated event bulletins (SEL3) and the corresponding analyst-reviewed bulletins, *late event bulletins* (LEB). Each bulletin includes data for both events and arrivals, as well as associated information. The parametric data used to train the classifiers consists of a long list of attributes related to the arrivals and associations. This list is detailed in Table 1.

Table 1. List of attributes used to form the feature vector to characterize SEL3 events.

Origin table fields: lat, lon, depth, time, nass, ndef, ndp, grn, srn (vectorized), dtype (vectorized), mb, ms, ml
Origerr table fields: sxx, syy, szz, stt, sxy, sxz, syz, stx, sty, stz, sdobs,
smajax, sminax, strike, sdepth, stime
Number of associated arrivals at each station
Haversine distance from origin location to each station
Counts of associated Assoc.delta values in six-degree bins
Counts of associated Arrival.qual values
Counts of associated Assoc.phase values
Counts of associated Arrival.iphase values
Number of times Assoc.phase != Arrival.iphase for associated arrivals
Number of time defining associated arrivals
Fraction of time defining associated arrivals
Number of azimuth defining associated arrivals
Fraction of azimuth defining associated arrivals
Number of slowness defining associated arrivals
Fraction of slowness defining associated arrivals

Mean of absolute value of associated Assoc.timeres values
Mean of absolute value of associated Assoc.azres values
Mean of absolute value of associated Assoc.slores values
Mean of associated Arrival.snr values
Variance of associated Arrival.snr values
Mean of associated Arrival.deltim value
Variance of associated Arrival.deltim values
Mean of associated Arrival.delaz value
Variance of associated Arrival.delaz values
Mean of associated Arrival.delslo value
Variance of associated Arrival.delslo values

The three-month dataset used in the study consisted of IDC parametric data for mid-March to mid-June 2009. It included 13,254 SEL3 events with 150,275 associated arrivals. The LEB bulletin for that time period included 9,961 events with 169,981 associated arrivals. Of these LEB arrivals, 114,464 were automatically generated (i.e., retained by analysts from SEL3) and 55,517 were added by analysts.

A set of SEL3 events for use by the classification procedures is obtained by featuring numeric values that informatively summarize the parametric data describing each SEL3 event, including residuals, error ellipses, number of defining phases, signal-to-noise ratios, and other quantities (see Table 1 for the list of features). The set of such values for a given SEL3 event forms a real vector, with which a binary label is associated indicating whether or not that event was subsequently discarded by analysts. For evaluation, a data set is built that excludes parametric data based on non-seismic stations. The first 75% of SEL3 events are used for training, and the final 25% of SEL3 events are used as a test set. To evaluate a given classification procedure, it is first trained to learn a classifier. Subsequently, the classifier is evaluated to determine which previously unseen events will be discarded by analysts by applying it to the test set.

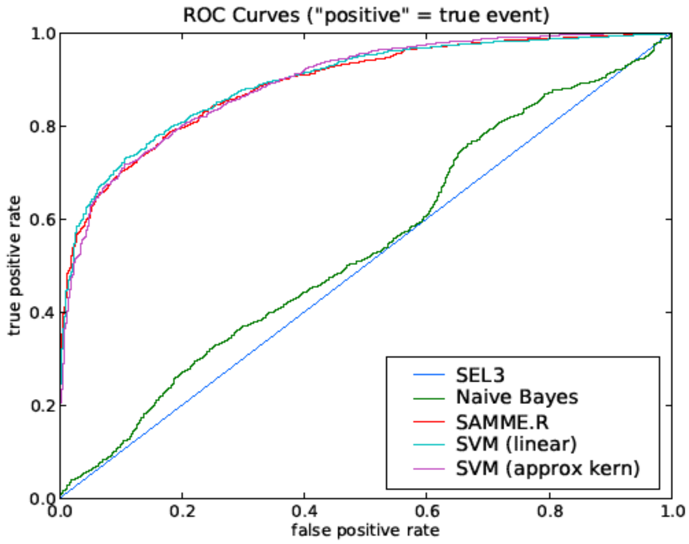

Figure 2. ROC for the different false events classifiers tested in this study (after Kleiner et al., 2009).

The primary metric used to assess the performance of the classification procedures is *accuracy*: the fraction of events in the test set that are correctly classified. The receiver operating characteristic (ROC) curves measure the tradeoff between false positive and false negative rates. Figure 2 summarizes the performance metrics for the different classifiers tested along with the SEL3 bulletin performance. The accuracy of the SEL3 bulletin under this measure is 65.41%. The classification procedures yield significant improvements. Among them, the linear L2 SVM yields the best performance with 81.47% accuracy. Other methods considered, such as boosted decision stumps and the non-linear L2 SVM, achieve similar performance, though the linear SVM enjoys the additional advantages of efficiency and interpretability. It is also worth noting that the simpler Naive Bayes procedure provides some improvement over SEL3 but significantly underperforms its more sophisticated counterparts.

Seismic Phase Identification

The second short-term project presented here is the identification of seismic phases on three-component seismic stations using features extracted from the waveform. This project aims at assessing potential improvements in the current operational system at the level of station processing, as shown on Figure 1 (Phase_ID label). A data set consisting of 117,181 detections at 3-component stations during 2009 was compiled from a number of IDC tables. The attributes used to perform the classification and their definitions are shown in Table 2. The ground truth phase for these detections is taken from the analyst-reviewed association table. For this study, only 8 phases were considered (Lg, P, PKP, Pg, Pn, Rg, S, Sn). In most cases, P and PKP are grouped together and the problem treated is a 7-class problem. One of the difficulties with this problem is that the P and PKP phases overwhelmingly dominate the statistics in terms of their numbers (75.7% of the records) and that other phases—Rg in particular—are very rare.

ddet60	Number of detections in a 60s window before and after the current detection
dtime60	Average time between detections in a 60s window before and after the current detection
Hmxmn	Maximum to minimum horizontal ratio
htov0.25	Horizontal to vertical ratio at 0.25 Hz
htov0.5 H	orizontal to vertical ratio at 0.5 Hz
htov1 H	orizontal to vertical ratio at 1.0 Hz
htov2 H	orizontal to vertical ratio at 2.0 Hz
htov4 H	orizontal to vertical ratio at 4.0 Hz
Hvrat	S-phase horizontal to vertical ratio
Hvratp	P-phase horizontal to vertical ratio
inang1	Long-axis incidence angle
inang3	Short-axis incidence angle
Per Dom	inant period
Plans S-phase	planarity
Rect Rectilin	earity
Slow Sl	owness

Table 2. Attributes used in the classification of different seismic phases detected on three-component stations.

Several classifiers were assessed against the analyst ground-truth and they are included in the following list:

Support Vector Machines – SVM algorithms identify a linear decision boundary in the feature space between two classes. Among all possible linear decision boundaries, they choose the one that maximizes the margin between the two classes. The solution allows for a number of points to lie on the wrong side of the boundary and assesses a penalty on them. In order to create a non-linear decision boundary in the original space, a kernel is used to generate additional features from the original ones. The 7-class problem is handled by building 7 binary classifiers that separate each class from all the others. For each new sample, all 7 classifiers are queried and the class corresponding to the one giving the highest prediction is chosen. See Cristianini et al. (2000) for more details on SVMs. The implementation used for this test was the LIBLINEAR library available at http://www.csie.ntu.edu.tw/~cjlin/liblinear/

Decision Trees – These are trees where the root and the other internal nodes contain a test of one of the features, which specifies whether a data point belongs to the left or right sub-tree of that node. The training data points are stored at the leaves of the trees. In order to classify a new test sample, the tests are used to find the correct leaf for the sample and the majority class of the training samples in that leaf is predicted. Learning of the decision trees is done by a greedy information-gain based procedure. See Hastie et al. (2000) for more details on decision trees.

Bagging – Bagging takes multiple bootstrap samples of the training data and learns a separate model from each bootstrap sample. The final prediction is made by averaging or voting the predictions of each model. Bagging may be applied to any classifier, but is commonly used in conjunction with decision trees to handle problems with over-fitting, which is what we did in this study.

Boosting – Boosting is conceptually similar to bagging except that each new model is constructed by reweighting the training data points according to how much prediction error they have in the current set of models. We tested a version called AdaBoost.

The methods were tested using the first 60% (in time) of the data for training and the last 40% for testing. There were two modes of modeling. The "single-task" mode lumps all the data together and learns a single model to be used at all stations. The "multi-task" mode learns a station-specific model using only the data for that station. A summary of the accuracy in predicting the phases for the test set is shown in Table 3.

Table 3. Accuracy results of testing six different classifying methods. The first column (single-task) shows the results of classifying independently of the station. The second column (multi-task) shows the results of classifying using station-specific classifiers.

Method	One predictor for all stations	One predictor per station
Multi-class SVM	76.09	82.32
Multi-class SVM with normalization	77.53	83.47
Real AdaBoost, 1vsAll	80.22	83.17
Gental AdaBoost, 1vsAll	80.31	83.42
Bagged decision trees	79.66	84.05
Bagged decision trees with kernels	**79.82**	**84.18**

The best performing models were those using bagged decision trees (shown in bold in Table 3). The kernel used to produce the best result was one that generates all quadratic features (i.e., the product of all pairs of features). Based on this empirical study, a supervised classification method using bagged decision trees with additional features generated by a quadratic kernel yields the best results. The best performance is obtained by training a separate classifier for each station.

Vertically Integrated Seismic Association

The goal of this project is to investigate if improvements can be made to the existing IDC seismic processing system using modern methods based on Bayesian inference. In a first stage, as indicated in Figure 1, the method is applied at the stage where parametric detection data is processed to obtain automatic bulletins. This initial project leads to the building of the NET-VISA prototype and the current results are presented here. A second stage, SIG-VISA, will include the detection and phase identification stage leading to the parametric data into the inference construct.

The probabilistic model in NET-VISA is based on a generative forward model which consists of:

- Spatial, temporal, and magnitude probabilistic models for event generation. The temporal model is a Poisson process. The spatial model is based on a prior spatial distribution of seismicity (shown in Figure 3) to which is added a uniform spatial distribution to take into account the possibility of an event occurring at a place where it never occurred before, as in the case of a newly developed nuclear test site. The magnitude model is based on the Gutenberg-Richter distribution.

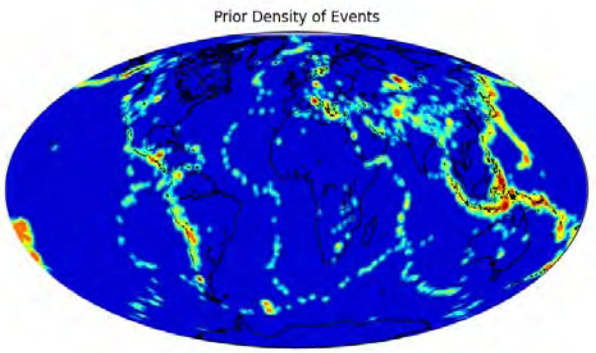

Figure 3. Prior distribution of events used in the event distribution generative model.

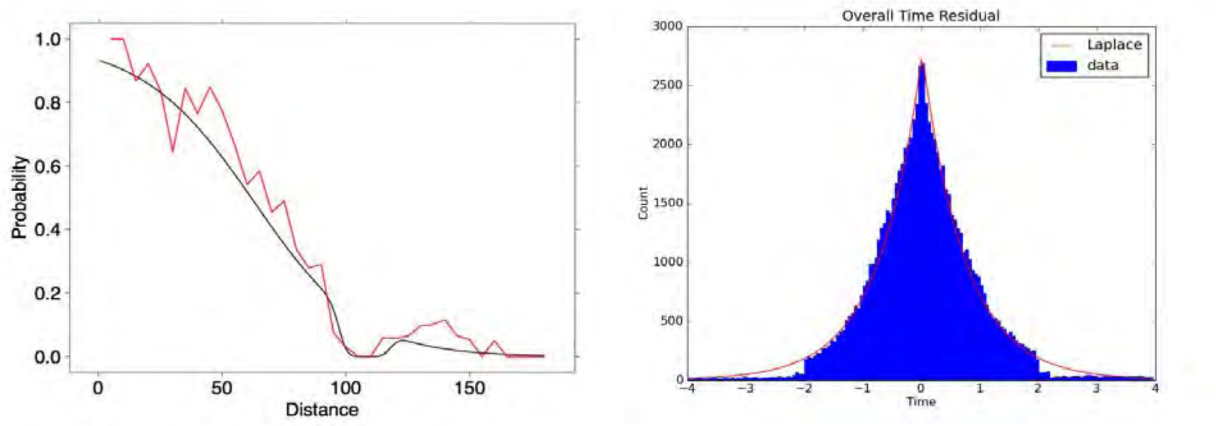

Figure 4. The curve to the left is the probability of detection of a P phase at station ASAR as a function of distance for an event of magnitude between 3 and 4. The analytical curve (in black) used in the calculations is shown superimposed on the observations (in red). The histogram to the right shows the overall (over all stations) travel time residuals distribution. Note the good fit to a Laplace distribution.

- Detection generation model at a given station knowing the location and magnitude of the event. This includes the phase-dependent probability that the event be detected at the station and the phase-dependent probability that the detection be of a certain amplitude and at a certain time around the mean time of arrival from the specific phase. Figure 4 shows on the left the analytical curve used in the probability of detection calculations and how it calibrates well against the observations. The right side of the figure shows the overall distribution of arrival time residual for P phases. Note that it fits very closely the Laplace distribution used in the modeling. Slowness and azimuth residuals are also modeled using a Laplace distribution.

- False detection. To take into account the fact that many false detections are made and may coincidentally be associated with an event, false detections are modeled into the generative forward model. The result is that the inference process takes into account the possibility of a false association. This is superior to the current operational model where any match between event and detection is assumed to be a positive match independently of the potential for the detection to be false. This leads to spurious association and degradation of location capability.

Using this probabilistic generative model, consisting of the product of all aforementioned probabilities, an inference engine is used to build event hypotheses consistent with the parametric observations. The general manner in which event set hypotheses are improved is that the current state is updated using a limited set of moves. The different moves used within the inference engine are listed below and give an intuitive view of the various steps leading to incremental improvements in the set of event hypotheses.

1. **Birth Move**: The birth move is executed only once in each event window. To propose the birth move events are searched on a fixed grid of points. The grid consists in 1-degree longitude and latitude buckets, 100-km depth buckets, 5-second time buckets within the event window, and 0.5-magnitude steps. At each of these grid points an event is hypothesized and the best detections added to it. At each station, the best detection with a score > 1 is added. Finally, the score of all the possible events is computed and the event with the greatest score > 1 is picked. If an event is found, then it is added to a list and all of its associated detections are marked unavailable. The algorithm is repeated over the same grid of events but with fewer associated detections. Finally, when no more events are found, the list of events is added to the hypothesis. The events are added to the hypothesis without any associated detections. Therefore this move is technically a downhill move. However, it is followed by other moves which will either add detections and make these events viable or kill them.

2. **Improve Detections Move**: For each detection in the detection window, all possible events up to the time of the detection are considered as well as all possible phases for these events to find the best event-phase for it. If the best event-phase has a score < 1 or the best event-phase already has another detection with a higher score then this detection is changed to a false detection.

3. **Death Move**: Any event with a score < 1 is killed and all its detections are marked as noise.

4. **Improve Events Move**: For each event we look for 1000 points chosen uniformly at random in a small ball around the event (5 degrees in longitude and latitude, 200 km in depth, 50 seconds in time, and 1 units of magnitude) for an event with a higher score. If such an event is found, the event parameters are changed.

5. **Final Pruning**: Before outputting events a final round of pruning to remove some duplicate events is performed. Any event which has another event with a higher score and within 5 degree distance and 50 second time is pruned. The reason for these spurious events is that true events have a long waveform coda within which detections are made at each site. These detections don't correspond to any event-phase of the original event but taken together they do suggest a new event at about the same location and time as the original event. Our model does allow for false detections but not false detections which are associated with a true event. Thus we need to explicitly prune these events for now. A later, more realistic model might be envisioned where the most probable detections within the coda are modeled.

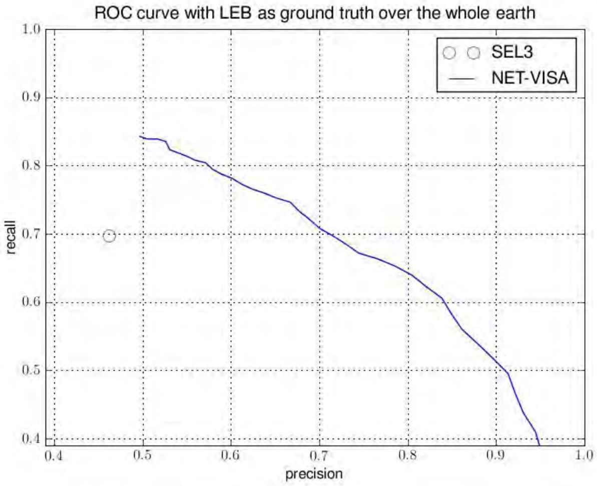

Figure 5. Receiver Operating Characteristics (ROC) for NET-VISA (in blue) and the location on the SEL3 bulletin in the precision/recall space.

The current results of the prototype are very promising when compared to the IDC operational results (see Figure 5). NET-VISA has a 14% higher recall (obtained events) at a slightly higher precision (automatic event that match an LEB event) than global association (GA), the current network processing program, and a 24% higher precision rate at the same recall rate as GA. A full assessment of the prototype cannot be achieved without analyst review of its results. There is tantalizing suggestion that this may prove very interesting since some of the events formed by NET-VISA (and missed by LEB) have been confirmed by bulletins (National Earthquake Information Center) independent of the IDC bulletin. It should be noted that an LEB bulletin based on the results of NET-VISA has the potential to be quite different especially at magnitudes between 3 and 4, and perhaps more accurate than the currently produced LEB which takes as starting point for analysis the SEL3 produced by GA.

CONCLUSIONS AND RECOMMENDATIONS

The few projects initiated in the Machine Learning area at the IDC of the CTBTO have proven that bringing innovative components to the current operating system should strengthen it and increase the precision of the automatic results which in turn should greatly increase the productivity of analysis and allow more precise bulletins to be produced. Some of the methods can be applied to some extent as a tuning exercise of the existing system, since it was found that some projects presented in this paper would benefit from station-dependent adjustment in parameterization of the algorithms. The NET-VISA project has shown the benefit of a radical overhaul of a key component of the system, and how new paradigms can be applied operationally with the increased computing power available to us over a decade after the development of the current IDC operational system.

A crucial step before a complete proof-of-concept for these projects can be presented to the member states of the CTBTO is to involve seasoned analysts in the evaluation of these new algorithms.

Another conclusion of these positive results of the current machine learning efforts is that more problems of interest to the CTBTO could be tackled such as infrasound phase identification, event screening, and on-site-inspection data sifting.

ACKNOWLEDGEMENTS

We thank Dr. Lassina Zerbo, IDC Director, for allowing us to publish this research and for his support of the machine learning efforts at the IDC and Misrak Fisseha for her analyst expertise and efforts in evaluating the FEI software delivered at the IDC.

REFERENCES

Arora, N., Jordan, M., Russell, S., and Sudderth, E., 2009, Vertically Integrated Seismological Analysis I: Modeling

Arora, N., Jordan, M., Russell, S., and Sudderth, E., 2009, Vertically Integrated Seismological Analysis II: Inference

Cristianini N. and J. Shawe-Taylor (2000). An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). The elements of statistical learning: Data mining, inference, and prediction (Second Edition). New York: Springer Verlag.

Mackey, L.; Kleiner, A.; Jordan, M. I., 2009, Improved Automated Seismic Event Extraction Using Machine Learning**,** American Geophysical Union, Fall Meeting 2009, abstract #S31B-1714.

Schneider J., Given, J., Le Bras, R., and Fisseha, M., 2010, Supervised Classification Methods for Seismic Phase Identification, EGU abstract, EGU2010-6269

Tuma, M. and Igel, C., 2009, Kernel-based machine learning techniques for hydroacoustic signal classification, ISS09 Conference.

Disclaimer

The views expressed in this paper are those of the authors and do not necessarily reflect the views of the CTBTO Preparatory Commission.